**Bachelor of Science in**
**Electronics and Telecommunication Engineering**


# Investigation and Transforming Resource-Scarce Language Question Generation: A Study on Text-to-Text Transformer Models for Bengali

*by*

Fahad Siddique Faisal

ID: 1808046

Department of Electronics and Telecommunication Engineering
**Chittagong University of Engineering and Technology**
Chattogram-4349, Bangladesh.
March, 2024

**Bachelor of Science in**
**Electronics and Telecommunication Engineering**


# Investigation and Transforming Resource-Scarce Language Question Generation: A Study on Text-to-Text Transformer Models for Bengali

*by*

Fahad Siddique Faisal

ID: 1808046

This thesis is submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Electronics and Telecommunication Engineering

Department of Electronics and Telecommunication Engineering
**Chittagong University of Engineering and Technology**
Chattogram-4349, Bangladesh.
March, 2024

# Candidates's Declaration

I hereby declare that the work in this Thesis has not been previously submitted to meet the requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the Thesis contains no material previously published or written by another person except where due reference is cited. Furthermore, the Thesis complies with CUET'S PLAGIARISM and ACADEMIC INTEGRITY regulations.

Signature of the Candidate

---

**Fahad Siddique Faisal**
ID: 1808046

# Electronic Submission of MRP

I at this moment declare that I am the sole author of the MRP. This is a true copy of the MRP, including any final revisions, as accepted by my Examiners.

I hereby also declare that the work contained in this Thesis can be uploaded to the repository of the library, Chittagong University of Engineering and Technology 1 year from the day of submission.

I authorize Chittagong University of Engineering and Technology to land this MRP to other institutions or individuals for scholarly research.

I further authorize the Chittagong University of Engineering and Technology to reproduce this MRP by photocopying or by other means, in total or part, at the request of other institutions or individuals for scholarly research.

I understand that my MRP may be made electronically available to the public.

Signature of the Candidate

---

**Fahad Siddique Faisal**
ID: 1808046

# Approval by the Supervisor

This is to certify that FAHAD SIDDIQUE FAISAL has carried out this research work under my supervision and that he has fulfilled the relevant Academic Ordinance of the Chittagong University of Engineering and Technology so that he is qualified to submit the following Thesis in the application for the degree of BACHELOR of SCIENCE in ELECTRONICS AND TELECOMMUNICATION ENGINEERING. Furthermore, the Thesis compiles with CUET's PLAGIARISM and ACADEMIC INTEGRITY regulations.

Signature of the Supervisor

---

**Prof. Dr. Md. Azad Hossain**
Head
Department of Electronics and Telecommunication Engineering
Chittagong University of Engineering and Technology
Chittagong-4349

# Acknowledgment

I would like to express my sincere gratitude to my supervisor, Prof. Dr. Md. Azad Hossain, for his invaluable guidance, support, and encouragement throughout this thesis project. Their expertise and insightful feedback were instrumental in shaping this work.

I am also grateful to my teachers who laid the foundation for my academic journey. Their dedication to teaching instilled in me strong critical thinking skills and a thirst for knowledge.

I would like to thank my learned friends. Their willingness to share their knowledge and help me solve problems throughout this process proved to be a tremendous asset. This thesis would not have been possible without the support of these individuals. I am truly grateful for their contributions.

Last but not least, I want to thank me. I want to thank me for believing in me. I want to thank me for doing all this hard work. I want to thank me for having no days off. I want to thank me for for never quitting. I want to thank me for always being a giver and trying to give more than I receive. I want to thank me for trying to do more right than wrong. I want to thank me for just being me at all times.

# Abstract

The process of creating potential questions from text or other types of data, like documents, images, graphs, etc., is commonly referred to as Automatic Question Generation (AGQ) or simply Generating Question System. This intriguing subject has recently sparked a lot of attention in the field of natural language processing. The primary justification for this is the wide range of industries in which it can be used, including business, healthcare, and education, where it can be used for multiple choice and frequently asked questions inquiries. The majority of studies on question generation has been conducted in languages with abundant resources, like English. However, languages with fewer resources, like Bangla, were unable to advance significantly in this area.

In this work, we have considered this and have employed a series of refined text-to-text transformer (T5) based models.

This encoder and decoder model is a transformer-based BanglaT5 model along with Google's transformer based MT5-Base model. Investigating and Optimizing the models to achieve a high F-1 score in RougeL and BLEU-1 is the aim.

When the context of the text and the response are given, this unique model can be used to create inquiries that seem human.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1: INTRODUCTION

*This chapter provides an overview of the thesis. Section 1.1 discusses the research background, while section 1.2 outlines the problem statement that the research aims to address. Section 1.3 discusses the fundamental objectives of the research. Section 1.4 outlines the research's Significance. Lastly, section 1.5 provides an overview of the remaining chapters of the thesis.*

## 1.1 Research Background

A question serves as a linguistic tool to seek details, and a response often provides specifics about something. Question generation (QG) involves the generation of questions based on a given passage or context [5]. Due to its significance in optimizing question-answering systems by means of additional apps and data augmentation natural question creation is becoming more and more popular in both the academic and business sectors.[6] An essential component of the early adopters' historical usage of QG for teaching was the creation of close-ended or gap-fill questions.

## 1.2 Problem Statement

In being the 6th most frequently used tongue across the whole world, with 230 million native speakers, Bangla is lagging in the community of Natural language processing because of the scarcity of resources. In recent times, many NLP works have been done including the Bangla Natural Language Understanding Benchmark and Bangla Natural Language Generation Benchmark [7] which also includes Dialog generation, sentiment analysis [8], fake news detection and classification [9], summarization [7] and question answering [8]. A very negligible amount of work has been done in the field of question generation. It is known to all that, for answering a question it is essential to formulate the question properly. Sometimes, a well-crafted question can lead to a better outcome. In the field of AI, it is an opportunity to create a system to create a question Generation system for the Bangla Language.

## 1.3 Objectives

In response to the dearth of research on the creation of Bengali questions, this study presents cutting-edge multilingual and Bengali models designed to generate context-aware and answer-aware questions. These models can be easily fine-tuned for specific domains, including education and industry, to enhance their applicability.

Previously, there is one Bangla Question generation task has been done and it was done with rule-based approaches [8]. But in the previous time, there was just one transformer-based study on the generation of questions as a part of the IndicNGL benchmark [10].

## 1.4 Motivation

In the field of question Generation, there had been done a very little work. As a Bangla Native Speaker, we were lagging behind in the field of NLP.Various works happened in Question generation, but not in our Bangla language. So, the motivation for working on this topics are-

- Removing Insufficiency of mothods to develop Questions : With this, it is possible to enhance existing question-answering system.
- Improving Performance of the existing chatbots : Performance of the currently existing chatbots can be improved by this work.
- Data Augmentation : Enriched articles and even Text-book can be created with the help of this.

## 1.5 Study Significance

A question is a linguistic phrase for looking for information, and a response can include details. Generating Questions is the process of developing questions based on a certain passage or context. The generation of natural questions has lately drawn a lot of interest from academic and business circles because of its essential function in improving question-answering systems through enhancing data and achieving additional goals. Early pioneers in QG for educational purposes primarily focused on creating gap-fill or close-ended questions. However, contemporary QG research is advancing rapidly, with its application spanning various fields including education [11], industrial use cases like chatbots [12], conversational systems, and the healthcare sector[13]. It also finds relevance in the field of

education contexts like learning a language and getting ready for an exam. QG has expanded its scope in relation to other languages, involving environments with diverse codes and cross-lingual users [14], the generation of distractors for multiple-choice questions in Swedish [15], and transformer-driven question formulation in Turkish [16], and Arabic [17]. Additionally, it has been used to generate fact-based questions in Finnish [18] and fact-based questions in Chinese.

## 1.6 Thesis Layout

The remaining thesis is structured as follows. Section 2 provides Various terminologies and overview related to the work and gives a summary of a few existing works on natural language generation on Bengali and Question Generation on Non-English Languages. Section 3 discusses the task definition, techniques, hyperparameters, and architectures of the constituent modules of the proposed system. Section 4 reports the experimental findings and extensive error analysis of the models. Section 5 points out the prospects of future development with concluding remarks.

# Chapter 2: LITERATURE REVIEW

*This chapter is structured into two main sections. In the first section, section 2.1, The background studies that has been made are explained. First, in Subsection 2.1.1 word embedding and word2vec. Sequentially, Encoder-Decoder Architecture in subsection 2.1.2, Attentioin Mechanism in subsection 2.1.3 and finally, the transformer in subsection 2.1.4. In the second section, related works performed in the field of Question Generation has been mentioned. 2.2.1 subsection shows the overview of the Bangla Natural Language Generation and latter subsection, 2.2.2 subsection gives a overview about the Question Generation in non-english Languages*

## 2.1 Background Study

### 2.1.1 Word Embedding and Word2Vec

Word embedding is one of the most used methods for representing document vocabulary. Word embedding, which is a dense vector, can capture a word's context within a document or sentence, its semantic and syntactic similarities, as well as surrounding words, and other variables.

Word2Vec is a program for learning word embedding that makes use of a neural network model and a large text corpus. Each distinct word is represented by a dense vector. This vector was carefully selected so that the degree of semantic relatedness between two words is directly indicated by the cosine similarity between vectors.

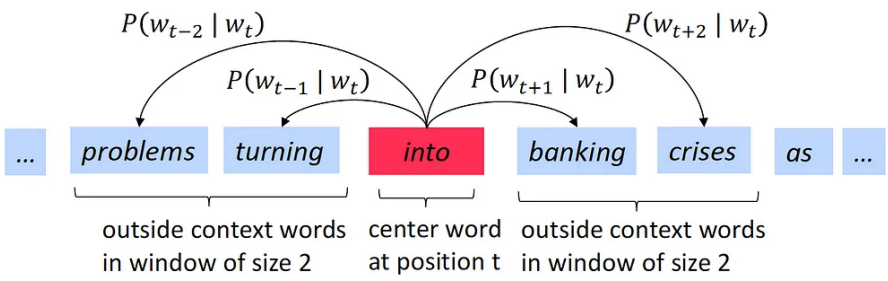


**Figure 2.1:** Association of a word with other surrounding words [1].

Cosine similarity between two word vectors determines the semantic relationship between two words. When the similarity score is 1 (or close) then the two vectors are similar, when 0 then two vectors are independent when -1 then two vectors would vary that when the similarity score is -1 then the words are similar but have opposite meanings, for example, words hot and cold.

### 2.1.2 Encoder-Decoder Architecture

Deep learning employs neural network design, namely Recurrent Neural Networks (RNN) and their variations, to address intricate issues. The encoder-decoder architecture, which was first introduced in 2014, is a widely used neural machine translation technique that can often get better results than traditional statistical machine translation methods. The system comprises three primary components: the encoder, responsible for transforming the input-sequence into a one-dimensional vector which is in fact is the hidden vector, and the output sequence is produced by the decoder, which transforms the hidden vector. The output sequence is produced by the decoder, which transforms the hidden vector into the sequence of a desired sequence. The models are trained together to optimize the conditional probabilities of the target sequence based on the input sequence. This architecture has been the core technology behind Google's translate service.

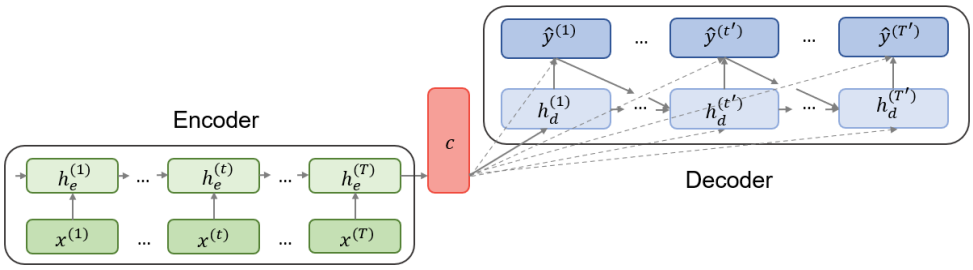


**Figure 2.2:** Encoder-decoder seq-to-seq model [2].

### 2.1.3 Attention Mechanism

When considering Deep Learning, the notion of attention is of the utmost importance in order to concentrate on criteria while processing data. This method draws inspiration from the human ability to focus on a certain activity. The Encoder-Decoder architecture in Deep Learning entails transmitting the ultimate hidden state of the encoder to the decoder, which might result in information loss as a result of compressing information into a vector of fixed size. Bi-directional layers are employed to tackle this problem by handling input sequences in reverse order, although their efficacy may be diminished for lengthier sequences.

The two main attention methods used in machine learning and natural language processing are Bahdanau Attention and Luong Attention. The Bahdanau Attention mechanism enables the decoder to consider the complete input sequence while decoding, resulting in improved exploitation of context. Luong Attention additionally prioritizes matching the decoder with segments of the input sequence, hence improving the overall effectiveness of the model.

Attention is a crucial factor in network architecture as it is responsible for overseeing and assessing the interconnection and reliance between different components. Deep learning models can enhance their ability to analyze and comprehend intricate sequences of data by integrating attention mechanisms like Bahdanau and Luong Attention. This integration results in enhanced performance in tasks such as machine translation and natural language processing.

#### 2.1.3.1 Bahdanau Attention

In 2014, Bahdanau et al. proposed attention to fix the information problem that the early encoder-decoder architecture faced commonly referred to as additive attention.[19] In the The Bahdanau attention mechanism involves obtaining of understanding in order to establish match between the input sequence and output sequence. This procedure requires concentrating on particular segments throughout each step of output sequence development. Alignment scores are computed by comparing the concealed states of the decoder and encoder through a feed-forward neural network. The softmax function is used to normalize the results and produce attention weights. The weights are subsequently merged with the hidden states of the encoder to generate a context vector. This context vector is then joined with the decoder input at each current time step.

#### 2.1.3.2 Luong Attention

Luong et al. proposed the second type of Attention. It is also known as Multiplicative Attention and was created on top of the Additive Attention mechanism.[20] The following are the two primary distinctions between Luong Attention and Bahdanau Attention:

- The procedure for calculating the alignment score.
- The decoder location where the Attention mechanism is introduced.

Luong's paper proposed three different types of alignment scoring functions as opposed to Bahdanau's attention method, which only used one type. Bahdanau's method computed a variable-length context vector that was then used as input

for the decoder. This context vector was also used to calculate the last decoder hidden state, hs−1. On the other hand, Luong et al.[20] calculated their context vector using the current decoder hidden state, hs, and used it to modify the decoder output before it was fed into the final softmax layer. This approach allowed for the creation of multiple scoring functions using the same attention mechanism.

### 2.1.4 Transformer

Vaswani[21] and Devlin[22] proposed the transformer model, which is a novel type of model that use self-attention to simultaneously evaluate all input words and capture the interdependencies among words inside a phrase. Transformers has a unique structure comprising of 6 layers, each containing 2 sublayers. This architecture enables them to efficiently capture and represent long-range connections, surpassing the capabilities of models based on Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). As a result, Devlin et al. (2018) created Bidirectional Encoder Representations [22] from Transformers, which are now the most advanced models utilized for transfer learning in Natural Language Processing (NLP).

The Transformer is a sophisticated deep learning model used in Natural Language Processing (NLP) to handle sequence-to-sequence tasks by effectively dealing with long-range connections between elements.

Transformers are neural networks that calculate representations of input and output without the need for convolution or sequence-aligned RNNs.

Contrary to earlier models such as RNNs and LSTMs, Transformers can process words simultaneously, resulting in improved speed and efficiency in capturing the context of language.

Transformers operate by enabling the model to examine additional words inside the input sentence, so facilitating contextual comprehension. Positional encodings are utilized to preserve the order of words, while their encoder-decoder architecture enables them to handle input text and produce output text. Transformers are frequently subjected to pre-training using a vast collection of text and subsequently adjusted to optimize their performance in specific tasks such as translation, summarization, or question answering. These models, such as BERT, GPT, DistilBERT, BART, and T5, have emerged and gained popularity. Each model possesses distinct features and can be used to different natural language processing (NLP) problems.
And in this research we are going to work with Transformer's T5 model.

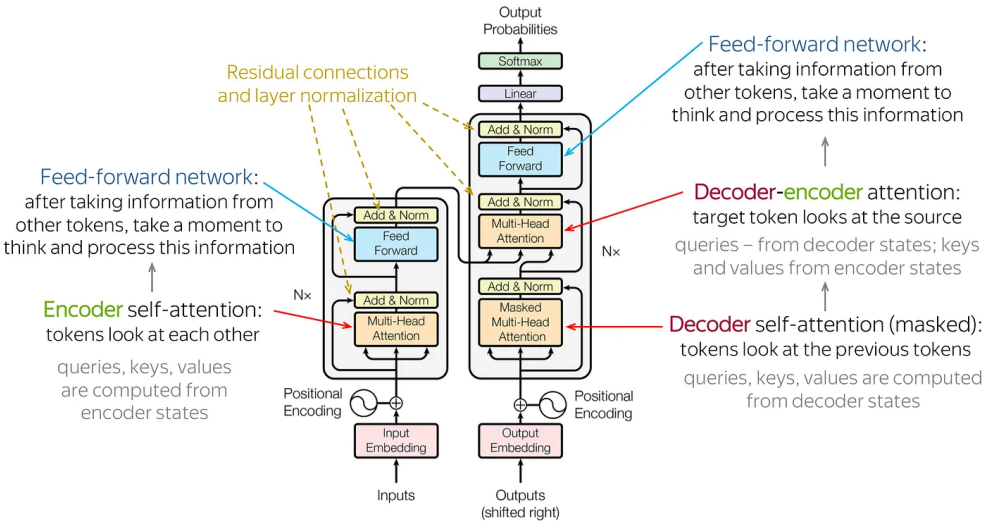


**Figure 2.3:** High-level overview of the Structure of Transformer [3].

## 2.2 Related Works

Rus et al.'s 2010 publication, "The First Question Generation Shared Task Evaluation Challenge," presented the first shared challenge for QG: generate questions from reading comprehension passages using linguistic elements and templates. Heilman and Smith's 2010 paper, "Good Question! Statistical Ranking for Question Generation," described a statistical ranking approach for generating questions from sentences based on features such as question type, word utility, and language model scores, allowing for the ranking and selection of the best questions.

### 2.2.1 Bangla Natural Language Generation

Natural language generation (NLG) is seeing a spike in attention due to the popularity of transformer models like OpenAI's GPT-3 [23]. NLG systems produce natural language text for use in writing, storytelling, screenplays for movies, poetry, dialogue, and answering questions in order to accomplish communication objectives [24]. Well-known AI support tools are Github's Copilot [25] and OpenAI's GPT-3. These systems can alter the persona, tone, topic, and content of created text by using controlled text generation. However, NLG systems get more complicated due to issues like hallucinations [26], picking the right evaluation criteria [24, 27], and erroneous or nonfactual information. Large-scale datasets for machine translation and abstractive text summarization [28] have been made available, but there is still much to learn about the wider use of transformer models for tasks like text generation, question or conversation creation, and text conversion. The "question generation" (QG) challenge, which

attempts to automatically produce questions based on a provided background paragraph, is the main subject of this work. While the IndicNLG benchmark [29] has investigated Bengali question production among other tasks, the Bangla NLG benchmark [30] covers tasks like machine translation, question answering, dialogue generation, and cross-lingual summarization.

### 2.2.2 Generating Question In Non-English Languages

Numerous researchers have shown interest in the topic of answering question, especially in languages other than English. To train question-answering systems, Mayeesha et al. translated SQUAD 2.0 into Bengali, and Bhattacharjee et al. benchmarked the BanglaBERT model. On the other hand, the problem of question generation has not been tackled. Question generation in other languages has been attempted: factual question generation [18], factoid-based question generation [31], distractor generation[15], transformer-based question generation in Turkish [32], and cross-lingual and code-mixed situations[33]. Answer-aware QG tasks have been implemented in English utilizing many T5 model iterations. To obtain the most accurate findings for non-English languages, T5 and other models have been used. Factoid question generation in Finnish has been achieved by the application of BERT and GPT-2 based models, with the greatest Rouge-L value of 0.33 [31].

**Table 2.1:** Related research of question generation in different languages

| Research Purpose | Dataset | Language | Models | Findings |
|---|---|---|---|---|
| TQG & QG [31] | SoqaQ-DynFi | Finnish | BERT-BASE, MuRIL, BERT-HugG, GPT-2-QG, GPT-2-HugG | The BERT-HugG dataset has the highest RougeL score of 0.53, surpassing other BLEU-n metrics and the METEOR score. |
| AQG from Swedish Documents [34] | SweQUAD-MC | Swedish | Fine-tuned BERT | Quantitative and qualitative investigation. |
| QG and answering [32] | TQuaD | Turkish | MT5-small, MT5-base, MT5-large | The study discovered that the TQuaD's fine tuning settings resulted in a score of 49.8 BLEU and 55.2 ROUGL1-score, whereas the same dataset also produced scores of 49.1 BLEU-1 and 54.3 ROUGL1. |
| Question generation [35] | mARCO | Arabic | TextRank Algorithm | 19.12 BLEU, 23.00 METEOR and 51.99 ROUGL-L |
| Answer aware question generation from Tabular text [36] | ToTTo | English | - | BLEU,ROUGL-1,ROUGL-L and METEOR metric for evaluating T5 model |
| AGS of Frequently Asked Questions [37] | Document related to COVID-19 | English | T5-base-qg-hl, T5-small-qg-hl | 21.3226 value BLEU, 43.5962 ROUGL-L, 27.0854 METEOR considering T5 models. |
| IndiGNLG [29] | SquAD | 12 languages with English | IndieBArT, mT5 | mT5 performed better in Generating question. |
| QG-Bench [38] | SQuAD v1.1 | 8 languages along with English | MT5-small, MT5-large, mBArT | English showed RougeL score 50.67, 52.95 is achieved in Japanese language |
| BQA | Sqaud.bn | Bengali | BanglaT5, MT5-small, MT5-base, mBArT | finetuned BanglaT5 generated 98 percent correct words gramatically and securing BLEU score 38.57 |

# Chapter 3: METHODOLOGY

*This chapter is structured to explain the procedure for developing the Question Generation and how to evaluate it. The steps involves selecting the dataset,pre-processing the data, development of the model, pre-training and fine-tuning the model and lastly using the Decoding Algorithms.*

The study focuses on the utilization of T5, a pre-trained English language model, to produce questions depending on the context and answer conditionally. The T5 model, utilizing a shared "text-to-text" structure for text-oriented natural language processing tasks, surpasses existing transformer-based models [39] in the domain of text production. The T5 model undergoes fine-tuning in the experiment, and the fine-tuned T5 model is then provided with sample data. The output prediction is evaluated using both human and automatic evaluation metrics. The T5 model underwent initial pretraining for language modeling tasks using a vast multilingual dataset, which encompassed the Bengali language as well. The model is trained using context words, which refer to the words that surround the masked words and are not themselves masked, to predict the masked words. The T5 model is optimized on a Bengali question-generating dataset that has been adapted from a question-answering dataset. The study illustrates that the passage in context and the response are given as the input in answer-aware question development.

## 3.1 Selecting data and dataset

There isn't a dataset available for the Bengali question generation. An existing question-answering dataset called BQA [40] was altered to handle question-generating tasks to solve this problem. This study made use of the training dataset that resulted from translating SQUAD 2.0. [41] The TidyQA secondary gold passage task served as the source for both the validation and test datasets. With 132777 samples overall, BQA is comparable to earlier studies in Arabic, Hindi, and Finnish. Using the BanglaT5 model, the dataset was utilized in the BanglaNLG benchmark to measure Bengali question answering. This methodology bears the resemblance to other research conducted on datasets and languages, including Arabic[35], Hindi [33], and Finnish[31].

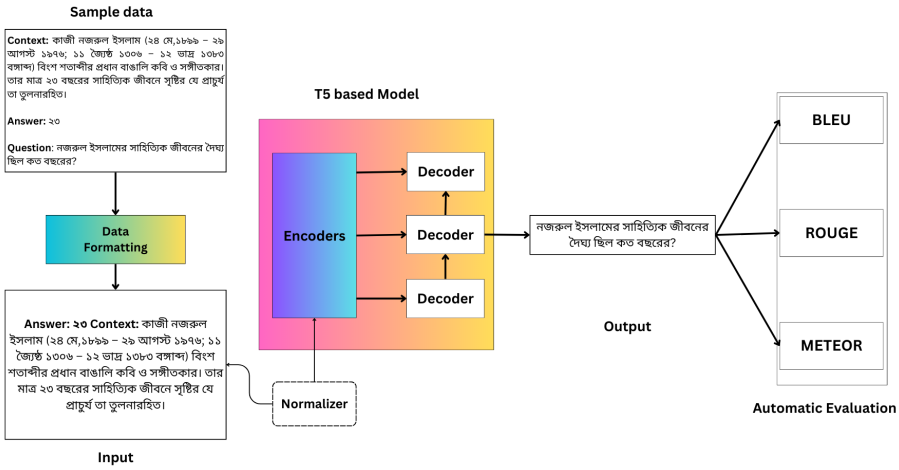


**Figure 3.1:** Simplified Methodology for Entire QG system

## 3.2 Preprocessing

Question generation research has been undertaken using two approaches: answer-agnostic and answer-aware. In the answer-agnostic[42] scenario where the answer is not considered, input passages are supplied without a provided answer. However, in this current scenario where the answer is considered, the answer to the question is already provided. This strategy is more straightforward to assess because it includes a reference question throughout the finetuning stage. Two ways are recommended for managing the many-to-one relation: All questions per line (AQPL) and One question per line (OQPL).

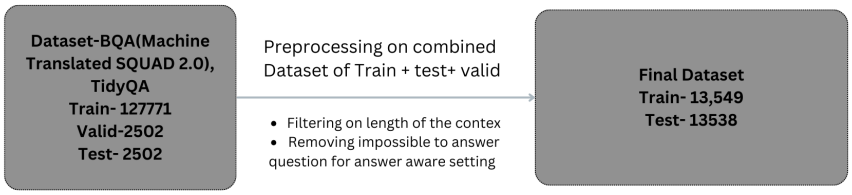


**Figure 3.2:** Dataset Statistics before and after the data-preprocessing

The trials are conducted within the framework of one specific question for

each passage. Following the application of a filtering process, a total of 13,549 samples are available for training purposes, whereas 13,538 samples are allocated for validation. The dataset has been pre-processed to meet the necessary format for generating questions that are aware of the answers. Only acceptable response pairs and filtered passages with a maximum context length of 500 characters have been retained. Unanswered questions are also excluded.

The data sample is provided in its original form, with the ”answer:” and ”context:” sections separated. The ”target” refers to a question sentence. The dataset contains a total of 132,777 samples, 20 percent of the data is then added to the validation set.

**Input :**
**answer:** মৌলভী আবদুস সামাদ
**context:**মতিউর রহমান ১৯৪১ সালের ২৯ অক্টোবর পুরান ঢাকার ১০৯, আগা সাদেক রোডের "মোবারক লজ"-এ জন্মগ্রহণ করেন। তার পিতার নাম মৌলভী আবদুস সামাদ এবং মাতার নাম সৈয়দা মোবারকুন্নেসা খাতুন।

**Target:**
মতিউর রহমানের বাবার নাম কি ?

**Figure 3.3:** Sample input state before and after the data processing

## 3.3 Models

The primary objective of this work is to generate questions using a modified dataset consisting of MT5-small, MT5-base, and BanglaT5 [40]. MT5 is a version of T5, which is a type of text-to-text transfer transformer that supports multiple languages and it has been trained using a dataset called mC4, which is derived from Common Crawl and encompasses 101 different languages. T5 is an English language model that is pre-trained and utilizes a fully packed ”text-to-text” structure to address all-natural language processing based on text (NLP) problems. Additionally, it treats classification problems and QA difficulties as text-based tasks, where labels like ”positive” or ”negative” are assigned to solve problems of sentence categorization.

IndicBART [43] is a model that is based on mBART and has acquired proficiency in several Indian languages, encompassing English. It consists of two

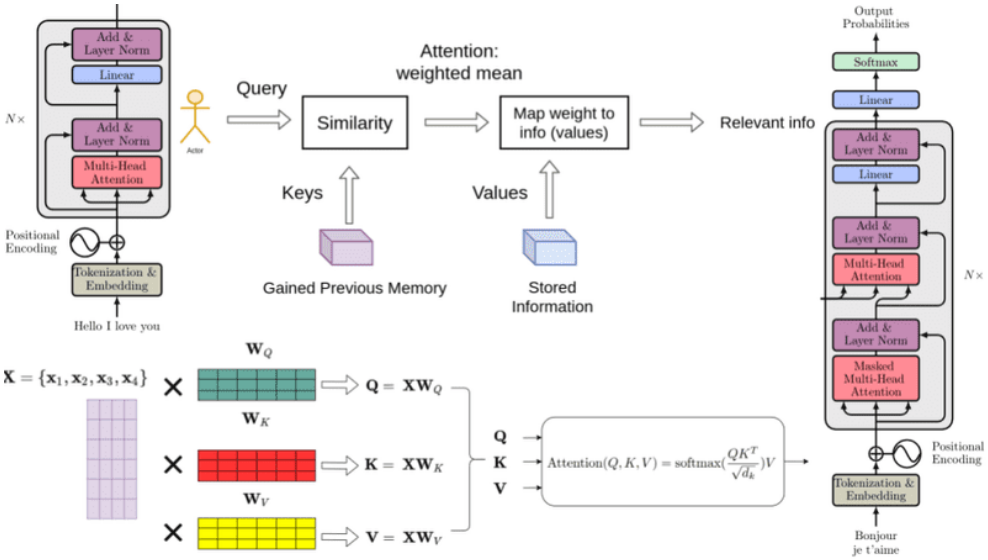


**Figure 3.4:** Text-To-Text Transfomer architecture [4]

variants, namely unified and Devanagari script. IndicBART utilizes the similarity in writing systems among Indic languages to improve the process of transfer learning. In general, these models present encouraging options for generating questions in multilingual environments.

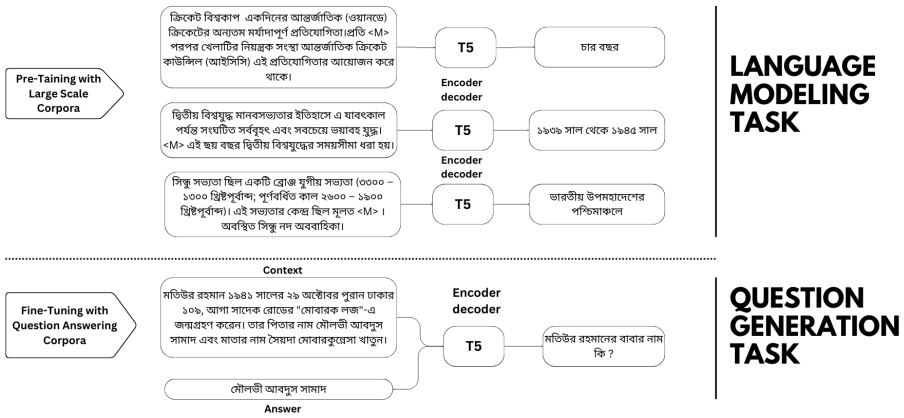


**Figure 3.5:** Overview of the Pretaining and finetuning for Bengali question Generation

In Fig.4.1,The T5 model, which was trained using a corpus of Bengali language data, is depicted in the upper half of the figure.<M> is indicating the masked tokenization. The model predicts masked tokens in input texts by employing Masked Language Modelling (MLM) [39].In MLM, a portion of the context is masked and the model is trained with contexual group of words and non-masked words which are the elements of the surrounding of the masking. Using a Ben-

gali question producing dataset that was transformed from a question answering dataset, the question generating finetuning stage is depicted in the lower part of the picture. Through this method, the T5 model for language modeling is improved. .

## 3.4 Decoding Algorithms

Decoding algorithms is the process of generating questions, to determine the subsequent word at each timestep. One of the decoding algorithms, Greedy decoding employs a strategy of selecting the word that has the most likelihood at each timestep, whereas Beam Search [44] methodically analyzes numerous potential options and ultimately selects the hypothesis having the largest cumulative likelihood. Top-k [45] and top-p [46] sampling are used in methods based on sampling, wherein a word is selected among the k most probable words at each time step. The values for top-k, as well as, top-p are set to 10, 50, and 90, with a range of 0.9 to 1. To prevent excessively long question generators, the maximum length has been limited to 50 tokens.In our research, we have propritize Beam-Search Algorithm for ensuring the imporved context and answer awareness since it analyzes the whole text for the context and select the best option from methodically generated various options.

## 3.5 Evaluation

BLEU[47], METEOR[48], ROUGE[49], and other automatic metrics are categorized under the first category in a recent study [50] on assessment metrics for natural language generation (NLG) systems. BLEU assesses the quality of text by computing n-grams between generated sentences and reference candidate texts. ROUGE, an algorithm focused on recall, offers automated tools for assessing the quality of sentences. METEOR adds a weighted F score to improve BLEU's recall and precision calculations.

# Chapter 4: EXPERIMENTS

*This chapter is structured to explain the procedure for various experiments that has been conducted. In section 4.1 explains the Implementation Settings that used in this research, section 4.2 explains the Evaluation Measures that has been used here. Section 4.3 explains the results and performance comparison that we obtained from the experiments.*

## 4.1 Implementation Settings

The notable platforms used in this experiment are:

- Google Collaborators, colloquially referred to as Colab, is a cloud-based platform developed by Google that allows users to write, execute, and distribute Python code right within their web browser. The platform provides a complimentary virtual space for conducting data analysis, machine learning, and educational activities. It is equipped with high-performance computer resources such as GPUs. The effortless incorporation of Google Drive streamlines the process of storing and distributing notebooks.

- GitHub: GitHub provides version control and collaborative tools for software development projects that use the Git revision management system. It offers an intuitive interface for managing repositories, allowing developers to track changes, collaborate, and manage codebase versions. GitHub also functions as a social coding platform, promoting collaboration, transparency, and efficient code management in the modern software development scene.

- Huggingface: Hugging Face, a leading machine learning startup, offers an open-source toolset for deploying and distributing pre-trained NLP models. The library has cutting-edge writing, translation, summarizing, and question-answering models. Hugging Face's platform improves teamwork and NLP. Enterprises can use the company's commercial services to develop and implement advanced NLP applications.

- Wandb: WandB simplifies machine learning model training, visualization,

and management. A central dashboard visualizes measurements and compares tests in real-time. WandB works with PyTorch, TensorFlow, and Keras, making it straightforward to incorporate. Since it automatically archives all experiment code, data, and hyperparameters, it prioritizes reproducibility and cooperation. This improves machine learning development efficiency and collaboration.

This experiment is carried out on the Google Collaboratory Platform, which is GPU-facilitated. Folium (0.9.1) is used for visualizing geospatial data, xl-sum is used for Multilingual Summarization, Unidic for segmentation, punkt for tokenizing text into sentences, and normalizer for reducing variation and inconsistencies in the preparation of the data.

## 4.2 Evaluation Measures

### 4.2.1 BLEU Score

Automatic machine translation text evaluation is done using the metric known as Bilingual Evaluation Understudy (BLEU) [**?**]. It is a corpus-level and popular statistic for assessing the value of text. The BLEU score, which ranges from 0 to 1, evaluates how closely machine-translated material resembles a group of superior reference texts. A number of 0 denotes that there is no overlap between the generated output and the reference text, whereas a value 1 denotes perfect overlap. Yet, even human translation does not get a score of 1.0. These equations are used to calculate BLEU scores:

$$\text{Geometric Average Precision}, N = \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log P_n\right) = \prod_{n=1}^{N} P_n^{\frac{1}{N}} \tag{4.1}$$

$P_n$ is the clipped precision for each n-gram where n = 1, 2, 3,...,N and $W_n$ is the uniform weights.

$$\text{Clipped Precision}, P_n = \frac{\text{clipped no of correct predicted words}}{\text{no of total predicted words}} \tag{4.2}$$

Brevity Penalty is used to penalize sentences that are too short.

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(\frac{1-r}{c}\right), & \text{if } c \leq r \end{cases} \tag{4.3}$$

Where c is the predicted sentence's word count and r is the target/reference's word count.

$$\text{BLEU(N)} = \text{Brevity penalty} \times \text{Geometric Average Precision Score(N)} \tag{4.4}$$

### 4.2.2 METEOR

The Metric for Evaluation of Translation with Explicit Ordering (METEOR) is built on the harmonic mean of unigram precision and recall, with recall weighted more heavily than precision [48]. The BLEU score has a number of flaws, including the fact that it ignores word order, does not verify that all terms in the reference are covered, does not account for semantic similarity, and only looks for exact word matches.

An alignment between the generated text and the reference can be done by match- ing word for word, or by using tools for similarity such as word embeddings, dic- tionary and so on. METEOR takes into account both the precision and recall while evaluating a match:

$$\text{precision}, p = \frac{m}{w_c} \tag{4.5}$$

$$\text{recall}, r = \frac{m}{w_r}, F_{\text{mean}} = \frac{10 \times p \times r}{r + 9p} \tag{4.6}$$

Where m denotes the number of unigrams in the proposed translation, which is also found in reference. The number of unigrams in the candidate translation is $w_c$. The number of unigrams in reference translation is $w_r$.

A chunk is a group of words that follow one another. We usually notice that chunks of words in the source correspond to chunks of words in the target. The number of chunks in the candidate that map to chunks in the target or reference determines the chunk penalty:

$$\text{chunk penalty}, p_c = 0.5 \times \left(\frac{c}{u_c}\right)^3 \tag{4.7}$$

Where c is the candidate's chunk count and $u_c$ is the candidate's unigram count. The F-score calculated from precision and recall along with the chunk penalty This final equation makes up the ultimate meteor score:

$$M = F_{\text{mean}}(1 - P) \tag{4.8}$$

## 4.3 Results

The study conducted a comparison between pre-trained models that are either monolingual or multilingual. The models included in the comparison were BanglaT5, IndicBART, and mT5, which is a model exclusively trained for Indic languages. The IndicBART model underwent training using parameters like those of MT5-base and BanglaT5. It outperformed sampling-based approaches in terms of ROUGE-L and BLEU scores.

**SAMPLE PREDICTION WITH BEAM SEARCH ALGORITHMS FOR PROPOSED METHOD**

| Reference Question | Predicted Question | Answer | Context |
|---|---|---|---|
| অ্যারিজোনার সীমান্তে মরুভূমির নাম কি? | মোজাভ মরুভূমির পূর্বে কোন মরুভূমি অবস্থিত? | কলোরাডো মরুভূমি | পূর্বে কলোরাডো মরুভূমি এবং কলোরাডো নদী অ্যারিজোনা সীমান্তে এবং নেভাদা রাজ্যের সীমান্তে মোজাভ মরুভূমি অবস্থিত। দক্ষিণে মেক্সিকো–যুক্তরাষ্ট্র সীমান্ত। |
| সময় এবং স্থান বা অনুরূপ পরিমাপের সীমা নির্ধারণ করতে প্রায়ই কি অ্যালগরিদম দ্বারা ব্যবহৃত হয়? | অ্যালগরিদম দ্বারা ব্যবহৃত সময় বা স্থানকে আবদ্ধ করে অনেক গুরুত্বপূর্ণ জটিল শ্রেণী কি নির্ধারণ করা যেতে পারে? | জটিল শ্রেণী | অ্যালগরিদম দ্বারা ব্যবহৃত সময় বা স্থানকে আবদ্ধ করে অনেক গুরুত্বপূর্ণ জটিল শ্রেণী নির্ধারণ করা যেতে পারে। এভাবে সংজ্ঞায়িত কিছু গুরুত্বপূর্ণ জটিল সিদ্ধান্ত সমস্যা নিম্নরূপ: |
| অরেঞ্জ, সান দিয়েগো, রিভারসাইড এবং সান বার্নার্ডিনো পাঁচটি কাউন্টির মধ্যে চারটি, সর্বশেষ কাউন্টিটির নাম কি? | অরেঞ্জ, সান দিয়েগো, সান বার্নার্ডিনো এবং রিভারসাইডের কাউন্টিগুলি কোথায় অবস্থিত? | লস এঞ্জেলেস | লস এঞ্জেলেস, অরেঞ্জ, সান দিয়েগো, সান বার্নার্ডিনো এবং রিভারসাইডের কাউন্টিগুলি রাজ্যের পাঁচটি সর্বাধিক জনবহুল এবং মার্কিন যুক্তরাষ্ট্রের শীর্ষ ১৫ টি জনবহুল কাউন্টিগুলির মধ্যে রয়েছে |

**Figure 4.1:** Sample Prediction with Beam Search Algorithm by the proposed method

Here in this experiment, our goal was to figure out the best possible scenario where the performance of the Bangla Question Generation System. In this experiment, two decoding algorithms are used. They are : Greedy algorithm and Beam algorithm. In case of BanglaT5 model, Greedy algorithm is performing better than any other decoding algorithm. But In case of the Mt5-base model, the Beam Algorithm is performing even better than BanglaT5 Beam algorithm in most of evaluation processes. So, In this current methodology, the model we proposed that the use of the MT5 model along with text normalizer with the Beam Search decoding Algorithm, which performs better generating Bangla Questions with better awareness to the context of the text.

### 4.3.1 Performance Comparison

IndicBART exhibited subpar performance across all measures when compared to T5 models that underwent fine-tuning. Therefore, it is not advisable to do addi-

**Table 4.1:** Evaluation of question generation using finally selected models.

| Model with Decoding algorithm | | Rouge1 | Rouge2 | RougeL | $\text{RougeL}_{\text{sum}}$ | SacreBLEU | Meteor |
|---|---|---|---|---|---|---|---|
| BanglaMT5 | beam_width = 3 | 29.56 | 12.54 | 17.29 | 26.345 | 11.45 | 0.20.66 |
| | beam_width = 5 | 35.0 | 16.4103 | 31.0 | 31.0 | 14.1418 | 0.2192 |
| | beam_width = 7 | 37.22 | 16.98 | 30.29 | 30.887 | 14.20 | 0.2167 |
| proposed model | beam_width = 3 | 35.88 | 16.20 | 36.006 | 36.236 | 9.480 | 0.2351 |
| | beam_width = 5 | 39.4646 | 17.6801 | 37.0202 | 36.4848 | 9.5636 | 0.2358 |
| | beam_width = 7 | 39.4646 | 17.832 | 36.72 | 36.167 | 9.5690 | 0.2456 |

tional research for error analysis and human review. According to the QG test, T5 and BART-based models frequently exhibited inferior performance in non-English languages as compared to English. Compared to the IndicNLG benchmark, the fine-tuned BanglaT5 model achieved a RougeL metric score of 37.22 , which is 6.64 points higher. By following our current methodology, MT5 model generated highest 39.4646 RougeL, which is 8.8846 points higher.

# Chapter 5: CONCLUSION

The work presents a novel Bengali question generation system that utilizes the encoder-decoder T5 model. This system is capable of generating a single interrogative question in an answer-aware style, taking into account the surrounding contexts. The Bengali T5 model and its multilingual variations were utilized, refined, and evaluated based on their relevance and grammatical accuracy in generating queries. The study examined the causes of low-quality question production and compared the findings with human evaluations. Nevertheless, the work's reliance on specific languages may not be effective in low-resource languages. The study also recognizes the possible danger of producing poisonous queries as a result of language models inheriting social biases. Future work intends to expand the system by incorporating new pre-trained models like as BLOOM and ChatGPT. It also wants to vary generation styles and enhance the system to enable comprehensive evaluation of generated questions using annotated data. The study also explores the capacity of automatically generated questions to generate a dataset for Bengali question answering.

## 5.1 Implications

### 5.1.1 Theoretical Implication

The study used synthetic datasets in Bengali, akin to the IndicNLG benchmark [29] and QG-bench [51], to compare multilingual and monolingual T5 and BERT models. However, the pretraining data (monolingual vs. multilingual), dataset (synthetic vs. human annotated), and pre-trained model selection all had a substantial impact on the performance. The findings imply that more investigation is required to examine the ramifications and domain adaption for Bengali. Like QG-bench tests, the T5-based models fared well in human evaluations, receiving high scores on criteria including grammatical precision, relevance to context passage, and answer.

### 5.1.2 Practical Implication

The study suggests a T5-based QG system for Bengali with the goal of producing a wide range of pertinent, grammatically sound, and well-written questions. Both automated and human measures are used to assess the system's performance. Low-resource languages like Bengali suffer from a dearth of NLP studies and datasets. Bengali Question Answering Systems can benefit from the open-source QG pipeline's ability to generate questions from textbooks, create FAQs for documentation or products, and curate discussion datasets for specific areas.

## 5.2 Limitations

There are still some limitations to this methodology. Primary limitations are :

- There is no existing dedicated dataset for the Question Generation field in Bangla. So, fine-tuning is quite hard and accuracy is not always maintained.
- Since there is no dedicated dataset, optimizing is hindered and there is a chance of happening of hallucination.
- With working with the T5 model, it is not possible to create a hybrid model with other various models because of the unique architecture of the T5 model.
- Whether the synonyms of the generated question gives the similar result or not is not checked in this experiment.
- This trained model can be inflienced by social convension and show result of toxic kind.[52]

## 5.3 Future Recommendations

- Improved Pre-training: Increases the quality and diversity of pre-training data, exposing the model to a broader set of factual knowledge and linguistic patterns.
- Fine-tuning with High-Quality Data: Uses high-quality, curated datasets tailored to the target task to produce more truthful and relevant results.
- Retrieval-Augmented Generation: Combines the T5 model with a retrieval component to ensure that generated outputs are factually accurate.

- Fact-checking and Consistency Scoring: Detects and eliminates hallucinated outputs that contradict known facts or lack internal consistency.
- Advanced Training: The T5 model is trained on adversarially created examples that contain hallucinations to recognize and avoid producing such outputs.
- Controlled Generation: Techniques like as prompt engineering, context conditioning, and learned attribute control are used to lead the T5 model to produce outputs that meet certain constraints or qualities.
- Human-in-the-Loop: Uses human feedback and supervision during the fine-tuning or generation process to detect and reduce hallucinations.
- Ensemble and Reranking: Combines numerous T5 models trained on various data or methodologies, then reranks the results based on factual consistency and believability.

# Bibliography


[1] S. Sharma, "Words are numbers: A quick introduction about word embedding," 2019. Accessed: April 15, 2024.

[2] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[3] A. Abaskohi, "Navigating transformers: A comprehensive exploration of encoder-only and decoder-only models, right," 2024.

[4] The AI Summer, "Transformer Neural Network," 2022. [Online; accessed 12 April 2024].

[5] W. L. Taylor, ""cloze procedure": A new tool for measuring readability," *Journalism quarterly*, vol. 30, no. 4, pp. 415–433, 1953.

[6] J. F. Ruma, T. T. Mayeesha, and R. M. Rahman, "Transformer based answer-aware bengali question generation," *International Journal of Cognitive Computing in Engineering*, vol. 4, pp. 314–326, 2023.

[7] A. Bhattacharjee, T. Hasan, W. U. Ahmad, K. Samin, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla," *arXiv preprint arXiv:2101.00204*, 2021.

[8] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on bengali social media comments using machine learning," *International journal of cognitive computing in engineering*, vol. 4, pp. 21–35, 2023.

[9] N. Tabassum, S. Aodhora, R. Akter, J. Hossain, S. Ahsan, and M. M. Hoque, "Punny_punctuators@ dravidianlangtech-eacl2024: Transformer-based approach for detection and classification of fake news in malayalam social media text," in *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pp. 180–186, 2024.

[10] A. Kumar, H. Shrotriya, P. Sahu, R. Dabre, R. Puduppully, A. Kunchukuttan, A. Mishra, M. M. Khapra, and P. Kumar, "Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages," *arXiv preprint arXiv:2203.05437*, 2022.

[11] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A systematic review of automatic question generation for educational purposes," *International Journal of Artificial Intelligence in Education*, vol. 30, pp. 121–204, 2020.

[12] M. K. Lee and H. Park, "Exploring factors influencing usage intention of chatbot-chatbot in financial service," *Journal of Korean Society for Quality Management*, vol. 47, no. 4, pp. 755–765, 2019.

[13] W. Cai, C. Zhang, S. Zhang, S. Ai, Y. Bai, J. Bao, B. Chen, N. Chang, H. Chen, L. Cheng, *et al.*, "The 2021 china report of the lancet countdown on health and climate change: seizing the window of opportunity," *The Lancet Public Health*, vol. 6, no. 12, pp. e932–e947, 2021.

[14] D. Gautam, P. Kodali, K. Gupta, A. Goel, M. Shrivastava, and P. Kumaraguru, "Comet: Towards code-mixed translation using parallel monolingual sentences," in *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pp. 47–55, 2021.

[15] D. Kalpakchi and J. Boye, "Bert-based distractor generation for swedish reading comprehension questions using a small-scale dataset," *arXiv preprint arXiv:2108.03973*, 2021.

[16] F. C. Akyon, D. Cavusoglu, C. Cengiz, S. O. Altinuc, and A. Temizel, "Automated question generation and question answering from turkish texts," *arXiv preprint arXiv:2111.06476*, 2021.

[17] N. Malik, S. Muttakin, E. Lopez-Quiroga, N. J. Watson, P. Fryer, S. Bakalis, and O. Gouseti, "Microstructure and reconstitution of freeze-dried gum arabic at a range of concentrations and primary drying temperatures," *Food hydrocolloids*, vol. 104, p. 105712, 2020.

[18] M. Liu, V. Rus, and L. Liu, "Automatic chinese factual question generation," *IEEE Transactions on Learning Technologies*, vol. 10, no. 2, pp. 194–204, 2016.

[19] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[20] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, C. L., R. N., K. A., A.-R. M., S. R., B. A., A., and C. Raffel, "Attention Is All You Need," 2017.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[23] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, and D. Amodei, "Language Models Are Few-Shot Learners," 2020. Publication Title: In Proceedings of the 34th International Conference on Neural Information Processing Systems.

[24] A. Gatt and E. Krahmer, "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation," 2018.

[25] A. Ziegler, "GitHub Copilot research recitation," 2021.

[26] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung *Survey of Hallucination in Natural Language Generation. ACM Comput. Surv*, vol. 55, no. 12, 2023.

[27] R. Perera and P. Nand, "Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature," *Computing and Informatics*, vol. 36, pp. 1–31, 2017.

[28] T. Hasan, A. Bhattacharjee, M. Islam, K. Samin, Y.-F. Li, Y.-B. Kang, M. Rahman, and R. Shahriyar, *XL-sum: Large-scale multilingual abstractive summarization for 44 languages*. ArXiv Preprint ArXiv: 2106.13822, 2021.

[29] A. Kumar, H. Shrotriya, P. Sahu, R. Dabre, R. Puduppully, A. Kunchukuttan, A. Mishra, M. Khapra, and P. Kumar, "IndicNLG Benchmark: Multilingual Datasets for Diverse NLG Tasks in Indic Languages," 2022. Publication Title: In Conference on Empirical Methods in Natural Language Processing.

[30] A. Bhattacharjee, T. Hasan, W. Ahmad, K. Samin, M. Islam, A. Iqbal, M. Rahman, and R. Shahriyar, "BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla," 2022.

[31] I. Kylliainen, "Neural Factoid Question Answering and Question Generation for Finnish," 2022.

[32] F. Akyon, D. Cavusoglu, C. Cengiz, S. Altinuc, and A. Temizel, *Automated question generation and question answering from Turkish texts using text-to-text transformers*. ArXiv Preprint ArXiv: 2111.06476, 2021.

[33] V. Kumar, N. Joshi, A. Mukherjee, G. Ramakrishnan, and P. Jyothi, *Cross-lingual training for automatic question generation*. ArXiv Preprint ArXiv: 1906.02525, 2019.

[34] D. Kalpakchi and J. Boye, "BERT-based distractor generation for Swedish reading comprehension questions using a small-scale dataset," 2021. Edition: ArXiv Preprint ArXiv.

[35] S. Alhashedi, N. Suaib, and A. Bakri, *Arabic Automatic Question Generation Using Transformer Model*. EasyChair, 2022.

[36] S. Pandraju and S. Mahalingam, "Answer-Aware Question Generation from Tabular and Textual Data using T5," *International Journal of Emerging Technologies in Learning*, vol. 16, no. 18, 2021.

[37] Z. Chen, "Automatic Generation System of Frequently Asked Questions Based on the T5 Model," *Academic Journal of Science and Technology*, vol. 2, no. 1, pp. 42–45, 2022.

[38] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, "Generative Language Models for Paragraph-Level Question Generation," 2022. Edition: ArXiv Preprint ArXiv Pages: 2210 03992.

[39] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res*, vol. 21, no. 140, pp. 1–67, 2020.

[40] A. Bhattacharjee, T. Hasan, W. Ahmad, and R. Shahriyar, *BanglaNLG: Benchmarks and Resources for Evaluating Low-Resource Natural Language Generation in Bangla*. ArXiv Preprint ArXiv: 2205.11081, 2022.

[41] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," 2018.

[42] L. Dugan, E. Miltsakaki, S. Upadhyay, E. Ginsberg, H. Gonzalez, D. Choi, C. Yuan, and C. Callison-Burch, *A feasibility study of answer-agnostic question generation for education*. ArXiv Preprint ArXiv: 2203.08685, 2022.

[43] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. Khapra, and P. Kumar, "IndicBART: A Pre-trained Model for Indic Natural Language Generation," 2022. Pages: 1849–1863 Publication Title: In Findings of the Association for Computational Linguistics: ACL 2022.

[44] R. Reddy, "Speech understanding systems: Summary of results of the five-year research effort at Carnegie-Mellon University," 1977.

[45] A. Fan, M. Lewis, and Y. Dauphin, "Hierarchical neural story generation," 2018. Edition: ArXiv Preprint ArXiv.

[46] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, *The curious case of neural text degeneration*. ArXiv Preprint ArXiv: 1904.09751, 2019.

[47] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," 2002. Pages: 311–318 Publication Title: In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.

[48] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," 2005. Pages: 65–72 Publication Title: In Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.

[49] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," 2004. Pages: 74–81 Publication Title: Text Summarization Branches Out,.

[50] A. Celikyilmaz, E. Clark, and J. Gao, "Evaluation of text generation: A survey," 2020. Edition: ArXiv Preprint ArXiv.

[51] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, "Generative language models for paragraph-level question generation," *arXiv preprint arXiv:2210.03992*, 2022.

[52] T. Schick, S. Udupa, and H. Schütze, "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1408–1424, 2021.